Research Report

Maximum Marginal Likelihood Estimation With an Expectation–Maximization Algorithm for Multigroup/Mixture Multidimensional Item Response Theory Models

ETS RR–19-35

Jianbin Fu

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Maximum Marginal Likelihood Estimation With an Expectation–Maximization Algorithm for Multigroup/Mixture Multidimensional Item Response Theory Models

Jianbin Fu

Educational Testing Service, Princeton, NJ

A maximum marginal likelihood estimation with an expectation–maximization algorithm has been developed for estimating multigroup or mixture multidimensional item response theory models using the generalized partial credit function, graded response function, and 3-parameter logistic function. The procedure includes the estimation of item parameters, attribute population distribution parameters, and test takers' attributes. All estimation functions and derivatives are provided. This procedure has been implemented in an R program. A simulation study has been conducted using this R program on various models related to the generalized partial credit function, and the result shows reasonable parameter recovery.

Keywords Multigroup IRT; mixture IRT; multidimensional IRT; maximum marginal likelihood estimation with EM

doi:10.1002/ets2.12272

Since Bock and Aitkin (1981) first applied the maximum marginal likelihood estimation with expectation–maximization algorithm to item response theory (IRT) models, MML-EM has been widely used for these models. The application of (MML-EM) MML-EM to IRT models has been discussed in, for example, Baker and Kim (2004, Chapter 6), Cai (2010a), Gibbons and Hedeker (1992), Glas (1992), Moustaki (2000), Muraki and Carlson (1995), Verhelst and Glas (1993), and von Davier and Yamamoto (2004). The MML-EM presented in this research report for general IRT models is the direct extension of the algorithm for a single group unidimensional generalized partial credit model (GPCM) discussed in von Davier and Yamamoto (2004). In particular, it extends the GPCM (Muraki, 1992), graded response model (GRM; Samejima, 1969), and three-parameter logistic model (3PL; Birnbaum, 1968) to accommodate multiple attributes and multiple observed or latent groups.

The extended IRT models allow analyses of data from an instrument that measures multiple attributes on multiple groups. An attribute refers to a skill or knowledge that is measured by an item. For example, a Grade 4 mathematics test may measure the following four skills: numerical representations and relationships, computations and algebraic representations, geometry and measurement, and data analysis and personal financial literacy. Attributes are assumed to be continuous. The groups could be either observed or latent. If observed, they are referred to as multigroup IRT models (Bock & Zimowski, 1997); if latent, they are called mixture models (Rost, 1991; Uebersax, 1999; von Davier & Yamamoto, 2004). The models are quite flexible at the group level given that a model is identified: (a) Each observed group is allowed to take a different set of items; (b) item parameters and the distribution parameters of continuous attributes can be fixed, free, or constrained to be equal across groups; and (c) attribute sets are allowed to differ across groups. For a group, the population distribution of attributes is assumed to follow a multivariate normal distribution that is approximated by multivariate Gauss–Hermite quadrature (Davis & Rabinowitz, 1984; Heiss & Winschel, 2008; Jaeckel, 2005).

Although all or some of the multigroup/mixture multidimensional IRT models mentioned previously can be estimated by programs for complicated IRT models, for example, MIRT (Haberman, 2013), mdltm (von Davier, 2008; von Davier & Xu, 2009), mirt (Chalmers, 2012), and flexMIRT (Cai, 2017), the full details of the MML-EM estimation have not previously been presented. The MIRT package employs log-linear modeling and implements the maximum marginal likelihood method with the stabilized Newton–Raphson algorithm, which is different from MML-EM. Cai (2010a) provided

Corresponding author: J. Fu, E-mail: jfu@ets.org

a description of the MML-EM procedure in the appendix; however, his coverage of the maximization step was brief: He only mentioned "standard Newton-type" maximization and referenced the derivatives. For the derivatives of item parameters, he cited Baker and Kim (2004); however, this book only contains derivatives for unidimensional IRT models. For mirt and mdltm, the description in the documentation of MML-EM implementation is limited, especially on the maximization step. It appears that in mirt, the first and second derivatives of estimated parameters are not calculated based on explicit analytic derivations, and existing maximization routines are used. In this report, the Newton–Raphson method and the formulas of the first and second derivatives of all estimated item parameters and the first derivatives of attribute population distribution parameters are presented.

The main purpose of this report is to present a detailed description of the MML-EM algorithm for general IRT models that many psychometricians can follow. In the following sections, the estimation method is first described including MML-EM estimation of item parameters and attribute population distribution parameters, and attribute estimation (scoring). This is followed by a parameter recovery study on variant models of GPCM using an R program (R Core Team, 2017) that implements the MML-EM estimation described previously. Finally, a summary of the report is provided.

Estimation of Item Parameters and Attribute Distribution Parameters With a Maximum Marginal Likelihood Estimation With an Expectation–Maximization Algorithm

First, let us set up the notation. There are a total of I items and the associated J continuous attributes. The relationship between items and attributes is defined by an item by attribute incidence matrix $\mathbf{Q}$ with an element $q_{ij} = 1$ indicating that item i ($i = 1$ to I) relates to attribute j ($j = 1$ to J) and $q_{ij} = 0$ indicating that item i does not relate to attribute j in any way. There are a total of N test takers who belong to G observed or latent groups, and each group has N_g ($g = 1$ to G) test takers. Let $\mathbf{G}$ be a vector denoting all test takers' group memberships. The attribute set associated with group g includes J_g attributes, and the set of attribute indexes (i.e., j) is represented by $\mathbf{j}_g$. Note that attribute sets could be different across observed groups but must be the same across latent groups. The set of attributes in $\mathbf{j}_g$, denoted as $\boldsymbol{\theta}_g$, are assumed to follow a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_g$ and a variance–covariance matrix $\boldsymbol{\Sigma}_g$. The multivariate normal distribution is approximated by J_g-variate Gauss–Hermite quadrature points with D quadrature points on each dimension, resulting in a D^{J_g}-grid with node vector $\boldsymbol{\omega}_{gl}$ and weight π_{gl} ($l = 1$ to L_g, where $L_g = D^{J_g}$; Davis & Rabinowitz, 1984). Each element in $\boldsymbol{\omega}_{gl}$ is denoted as w_{glj}, referring to the quadrature point of attribute j in node vector l in group g. In each group g, test takers respond to a set of I_g items, and the set of item indexes (i.e., i) is denoted by $\mathbf{I}_g$. Note that item sets could be different across observed groups but must be the same for latent groups. Item i has M_i score categories, s_{im} ($m = 1$ to M_i). Item category scores s_{im} can be any real numbers; usually the integers from 0 to $M_i - 1$ are used, for example, $s_{i1} = 0$, $s_{i2} = 1$, and $s_{i3} = 2$. In each group g, there are P_g item response patterns $\mathbf{X}_{gp}$ ($p = 1$ to P_g) with individual item responses x_{gpi}, $i \in \mathbf{I}_g$, and there are N_{gp} test takers having response pattern $\mathbf{X}_{gp}$ with associated attribute vector $\boldsymbol{\theta}_{gp}$. Subscript g may be dropped from $\mathbf{X}_{gp}$ if item response patterns are the same across groups, for example, in latent groups, and subscript $p$is dropped from x_{gpi} if item response patterns are not referenced. Let N_p denote the number of test takers having response pattern $\mathbf{X}_p$ across groups. The set of all item response patterns across groups is denoted as $\mathbf{X}$, and the set of all test takers' attribute vectors across groups is denoted as $\boldsymbol{\theta}$. The probability of an item response depends on a test taker's attribute parameters and the item's parameters, and conditional independence is assumed among item responses (responses are independent conditional on the J attributes). The item parameter vector for item i in group g is denoted as $\boldsymbol{\eta}_{gi}$, and the item parameter vectors for all items and attribute population distribution parameters in group g, and across all groups, are denoted as $\boldsymbol{\eta}_g$ and $\boldsymbol{\eta}$, respectively.

The EM algorithm (Bock & Aitkin, 1981; Dempster, Laird, & Rubin, 1977) involves an iterative process that repeatedly executes two steps: an E step and an M step. In the E step, the expectation of the complete data log-likelihood with respect to the posterior distribution of missing data is estimated, leading to a marginal log-likelihood of the observed data. For IRT models, the unobserved (missing) data are test takers' attribute vectors, $\boldsymbol{\theta}$, and/or latent group memberships, $\mathbf{G}$. In the M step, the marginal log-likelihood is maximized with respect to item parameters and attribute distribution parameters, if estimated.

The complete likelihood of general IRT models based on the previous setup is

$$L(\mathbf{X}, \boldsymbol{\theta}, \mathbf{G}|\boldsymbol{\eta}) = \prod_{g=1}^{G} \prod_{p=1}^{P_g} \prod_{i \in \mathbf{I}_g} p\left(x_{gpi}|\boldsymbol{\theta}_{gp}, \boldsymbol{\eta}_{gi}\right)^{N_{gp}},$$

which can be approximated as

$$L^{\text{appr}}(\mathbf{X}, \boldsymbol{\theta}, \mathbf{G}|\boldsymbol{\eta}) = \prod_{g=1}^{G}\prod_{l=1}^{L_g} p\left(\boldsymbol{\omega}_{gl}, g\right)^{n_{gl}} \prod_{i\in \mathbf{I}_g}\prod_{m=1}^{M_i} p\left(x_{gi} = s_{im}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right)^{n_{glis_{im}}},$$

where the superscript "appr" in $L^{\text{appr}}(\mathbf{X}, \boldsymbol{\theta}, \mathbf{G}|\,\boldsymbol{\eta})$ indicates the approximation of the likelihood; $p(\boldsymbol{\omega}_{gl}, g)$ is the joint probability of group membership g and the attribute vector l in group g; n_{gl} is the number of test takers having attribute vector $\boldsymbol{\omega}_{gl}$; $n_{glis_{im}}$ is the number of test takers having attribute vector $\boldsymbol{\omega}_{gl}$ and getting score s_{im} on item i; and $p(x_{gi} = s_{im}|\,\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi})$ is the probability of obtaining item score s_{im} conditional on attribute vector $\boldsymbol{\omega}_{gl}$ and item parameters $\boldsymbol{\eta}_{gi}$. Then, the complete log-likelihood is approximately

$$\log\left(L^{\text{appr}}\right) = \sum_{g=1}^{G}\sum_{l=1}^{L_g}\left[n_{gl}\log p\left(\boldsymbol{\omega}_{gl}, g\right) + \sum_{i\in \mathbf{I}_g}\sum_{m=1}^{M_i} n_{glis_{im}}\log p(x_{gi} = s_{im}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi})\right]. \tag{1}$$

The iterative process stops if the convergence criterion is met. Usually, the convergence criterion is the largest change in an item parameter estimate and/or the ratio of the change of the approximate observed log-likelihood $[\log\left(L_{\text{obs}}^{\text{appr}}\right)]$ between two consecutive iterations over the previous $\log\left(L_{\text{obs}}^{\text{appr}}\right)$ being smaller than predefined values (e.g., .001 and .00001, respectively). The $\log\left(L_{\text{obs}}^{\text{appr}}\right)$ is estimated below.
For observed groups:

$$\log\left(L_{\text{obs}}^{\text{appr}}\right) = \sum_{g=1}^{G}\sum_{p=1}^{P_g} N_{gp}\log\left[\sum_{l=1}^{L_g}\pi_{gl}\prod_{i\in \mathbf{I}_g} p\left(x_{gpi}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right)\right],$$

for latent groups:

$$\log\left(L_{\text{obs}}^{\text{appr}}\right) = \sum_{p=1}^{P_g} N_p\log\left[\sum_{g=1}^{G}\sum_{l=1}^{L_g} p\left(\boldsymbol{\omega}_{gl}, g\right)\prod_{i\in \mathbf{I}_g} p(x_{gpi}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi})\right].$$

E Step

In the E step, the expectation of the complete data log-likelihood is estimated with respect to the posterior distribution of attributes, that is,

$$\log L(\mathbf{X}, \mathbf{G}\mid\boldsymbol{\eta}) = \sum_{g=1}^{G}\sum_{p=1}^{P_g} N_{gp}\sum_{i\in \mathbf{I}_g}\left[\int_{\mathfrak{R}}\log p\left(x_{gpi}|\boldsymbol{\theta}_{gp}, \boldsymbol{\eta}_{gi}\right) f(\boldsymbol{\theta}_g|\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g)d\boldsymbol{\theta}_g\right], \tag{2}$$

where $f(\boldsymbol{\theta}_g|\,\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g)$ is the density function of the attribute vector, $\boldsymbol{\theta}_g$, conditional on $\mathbf{X}_{gp}$, g, and $\boldsymbol{\eta}_g$. As for the approximate complete log-likelihood (Equation 1), this process is essentially to estimate the expected values of n_{gl} and $n_{glis_{im}}$ conditional on $\mathbf{X}_{gp}$, $\mathbf{G}$, and $\boldsymbol{\eta}_g$. To do that, for observed groups, the posterior distribution of $\boldsymbol{\omega}_{gl}$ conditional on $\mathbf{X}_{gp}$, g, and $\boldsymbol{\eta}_g$ is needed:

$$p\left(\boldsymbol{\omega}_{gl}|\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g\right) = \frac{\pi_{gl}\prod_{i\in \mathbf{I}_g} p\left(x_{gpi}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right)}{\sum_{l'=1}^{L_g}\pi_{gl'}\prod_{i\in \mathbf{I}_g} p\left(x_{gpi}|\boldsymbol{\omega}_{gl'}, \boldsymbol{\eta}_{gi}\right)}. \tag{3}$$

Then,

$$\overline{n}_{gl} = \sum_{p=1}^{P_g} N_{gp}p\left(\boldsymbol{\omega}_{gl}|\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g\right), \tag{4}$$

and

$$\overline{n}_{glis_{im}} = \sum_{p=1}^{P_g} N_{gp}I\left(x_{gpi} = s_{im}\right)p\left(\boldsymbol{\omega}_{gl}|\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g\right), \tag{5}$$

where $\overline{n}_{gl}$ and $\overline{n}_{glis_{im}}$ are conditional expected values of n_{gl} and $n_{glis_{im}}$, respectively; $I(x_{gpi} = s_{im})$ is the indicator function with value 1 if $x_{gpi} = s_{im}$, and 0 otherwise. If group g is latent, then the joint posterior distribution of $\boldsymbol{\omega}_{gl}$ and g conditional on $\mathbf{X}_{gp}$ and $\boldsymbol{\eta}_g$ is needed:

$$p\left(\boldsymbol{\omega}_{gl}, g|\mathbf{X}_p, \boldsymbol{\eta}_g\right) = \frac{p\left(\boldsymbol{\omega}_{gl}, g\right)\prod_{i\in\mathrm{I}} p\left(x_{pi}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right)}{\sum_{g'=1}^{G}\sum_{l'=1}^{L_g} p\left(\boldsymbol{\omega}_{g'l'}, g'\right)\prod_{i\in\mathrm{I}} p\left(x_{pi}|\boldsymbol{\omega}_{g'l'}, \boldsymbol{\eta}_{g'i}\right)}. \tag{6}$$

Then,

$$\overline{n}_{gl} = \sum_{p=1}^{P_g} N_p p\left(\boldsymbol{\omega}_{gl}, g|\mathbf{X}_p, \boldsymbol{\eta}_g\right),$$

and

$$\overline{n}_{glis_{im}} = \sum_{p=1}^{P_g} N_p I\left(x_{pi} = s_{im}\right) p\left(\boldsymbol{\omega}_{gl}, g|\mathbf{X}_p, \boldsymbol{\eta}_g\right). \tag{7}$$

By replacing n_{gl} and $n_{glis_{im}}$ in Equation 1 with their conditional expected values, the approximate complete log-likelihood becomes the approximate marginal log-likelihood function,

$$\log\left(L_m^{\mathrm{appr}}\right) = \sum_{g=1}^{G}\sum_{l=1}^{L_g}\left[\overline{n}_{gl}\log p\left(\boldsymbol{\omega}_{gl}, g\right) + \sum_{i\in\mathrm{I}_g}\sum_{m=1}^{M_i}\overline{n}_{glis_{im}}\log p(x_{gi} = s_{im}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi})\right]. \tag{8}$$

If attribute distribution parameters are estimated, $\boldsymbol{\omega}_{gl}$ and π_{gl} at iteration $t+1$ are updated based on the estimated attribute distribution parameters from the M step at iteration t, and if attribute distribution parameters are fixed, $\boldsymbol{\omega}_{gl}$ and π_{gl} are fixed during estimation. The joint probability $p(\boldsymbol{\omega}_{gl}, g)$ at iteration $t+1$ is equal to $\pi_{gl}\sum_{l=1}^{L_g}\overline{n}_{gl}/N$ for latent groups and $\pi_{gl}N_g/N$ for observed groups, where $\overline{n}_{gl}$ is estimated from iteration t and π_{gl} is the current value at iteration $t+1$.

M Step

In the M step, the marginal log-likelihood function is maximized with respect to the distribution parameters of continuous attributes, if estimated, and item parameters. The maximization takes two steps: one with respect to the item parameters and another with respect to the distribution parameters (Glas, Wainer, & Bradlow, 2000). The maximum number of iterations within the M step can be specified, and the same convergence criteria used for the EM cycles can also be applied here.

Item Parameter Estimation

For item parameters, the maximization uses the Newton–Raphson method (Atkinson, 1989). That is, item i's parameters in iteration $t+1$ are updated based on

$$\boldsymbol{\eta}_{gi}^{t+1} = \boldsymbol{\eta}_{gi}^{t} - \mathrm{H}^{-1}\left(\boldsymbol{\eta}_{gi}^{t}\right)\mathrm{L}\left(\boldsymbol{\eta}_{gi}^{t}\right), \tag{9}$$

where $\mathbf{L}\left(\boldsymbol{\eta}_{gi}^{t}\right)$ is a vector containing the first derivative of the marginal log-likelihood function (Equation 2) with respect to each item parameter; $\mathbf{H}\left(\boldsymbol{\eta}_{gi}^{t}\right)$ is the Hessian matrix, which is the second derivative matrix of the marginal log-likelihood function with respect to item parameters. According to an integration rule called *differentiation under the integral sign* (Lang, 1997, pp. 337–339), which is applicable here, the derivatives of the marginal log-likelihood function can be approximated by the derivatives of the approximate marginal log-likelihood function (Equation 8). Therefore, in the maximization process, we actually calculate the first and second derivatives of the approximate marginal log-likelihood function with respect to item parameters within an item:

$$\frac{\partial\log\left(L_m^{\mathrm{appr}}\right)}{\partial\eta_{iv}} = \sum_{g=1}^{G}\sum_{l=1}^{L_g}\left[\sum_{i\in\mathrm{I}_g, m=1}^{M_i}\overline{n}_{glis_{im}}\frac{\partial\log p\left(x_{gi} = s_{im}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right)}{\partial\eta_{iv}}\right],$$

and

$$\frac{\partial^2 \log\left(L_m^{\text{appr}}\right)}{\partial\eta_{iv}\partial\eta_{iv'}} = \sum_{g=1}^{G}\sum_{l=1}^{L_g}\left[\sum_{i\in \mathrm{I}_g, m=1}^{M_i} \overline{n}_{glis_{im}} \frac{\partial^2 \log p\left(x_{gi}=s_{im}|\boldsymbol{\omega}_{gl},\boldsymbol{\eta}_{gi}\right)}{\partial\eta_{iv}\partial\eta_{iv'}}\right],$$

where η_{iv} and $\eta_{iv'}$ are item parameters of item i. Thus the calculation of the derivatives of the approximate marginal log-likelihood function is essentially calculating the derivatives of $p(x_{gi}=s_{im}|\,\boldsymbol{\omega}_{gl},\boldsymbol{\eta}_{gi})$ with respect to item parameters of item i. In the following, the derivatives for the multigroup and multidimensional response functions with the GPCM form (MGPCM), the GRM form (MGRM), and the 3PL form (M3PL) are presented.

Generalized Partial Credit Response Function

The response function of the MGPCM can be written as (von Davier, 2008)

$$\begin{aligned} p_{glis_{im}} &= p\left(x_{gi}=s_{im}|\boldsymbol{\omega}_{gl},\boldsymbol{\eta}_{gi}\right) = p\left(x_{gi}=s_{im}|\boldsymbol{\omega}_{gl},\mathbf{b}_i,\mathbf{a}_i\right) \\ &= \frac{\exp\left(b_{is_{im}}+\sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}s_{im}\right)}{\sum_{m'=1}^{M_i}\exp\left(b_{is_{im'}}+\sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}s_{im'}\right)}, \end{aligned} \tag{10}$$

where $b_{is_{i1}}+\sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}s_{i1}\equiv 0$; $b_{is_{im}}$ is the intercept parameter for score category s_{im}; a_{ij} is the discrimination (slope) parameter for attribute j; and w_{glj} is the attribute j in the attribute vector, $\boldsymbol{\omega}_{gl}$. The derivation of the derivatives of $\log\left(p_{glis_{im}}\right)$ with respect to $\mathbf{b}_i$ and $\mathbf{a}_i$ becomes easy by taking advantage of the properties of the exponential family. Equation 10 can be rewritten in the form of the exponential family as the following (Barndorff-Nielsen, 1978):

$$p_{glis_{im}} = \exp\left[b_{is_{im}}+\sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}s_{im} - A\left(\boldsymbol{\omega}_{gl},\mathbf{b}_i,\mathbf{a}_i\right)\right], \tag{11}$$

where $A\left(\boldsymbol{\omega}_{gl},\mathbf{b}_i,\mathbf{a}_i\right)=\log\left[\sum_{m'=1}^{M_i}\exp\left(b_{is_{im'}}+\sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}s_{im'}\right)\right]$. Then, for $q_{ij}=1$,

$$\frac{\partial \log\left(p_{glis_{im}}\right)}{\partial a_{ij}} = w_{glj}s_{im} - \frac{\partial A}{\partial a_{ij}},$$

and for $t\neq s_{i1}$,

$$\frac{\partial \log\left(p_{glis_{im}}\right)}{\partial b_{it}} = I\left(s_{im}=t\right) - \frac{\partial A}{\partial b_{it}}.$$

Note that the subscript s_{im} in $b_{is_{im}}$ is changed to t so as to distinguish it from item scores. The second derivatives are the following:

For $q_{ij}=1$ and $q_{ij'}=1$

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial a_{ij}\partial a_{ij'}} = -\frac{\partial^2 A}{\partial a_{ij}\partial a_{ij'}},$$

for $t\neq s_{i1}$ and $t'\neq s_{i1}$

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial b_{it}\partial b_{it'}} = -\frac{\partial^2 A}{\partial b_{it}\partial b_{it'}},$$

for $q_{ij}=1$ and $t\neq s_{i1}$

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial a_{ij}\partial b_{it}} = -\frac{\partial^2 A}{\partial a_{ij}\partial b_{it}}.$$

Based on a property of the exponential family (Barndorff-Nielsen, 1978),

$$\frac{\partial A}{\partial \eta_{iv}} = E\left(T_{iv}\right)$$

$$\frac{\partial^2 A}{\partial \eta_{iv} \partial \eta_{iv'}} = \operatorname{cov}\left(T_{iv}, T_{iv'}\right) = E\left(T_{iv} T_{iv'}\right) - E\left(T_{iv}\right) E\left(T_{iv'}\right),$$

where T_{iv} and $T_{iv'}$ are sufficient statistics for item parameters η_{iv} and $\eta_{iv'}$, respectively, and the expectations are with respect to x_{gi} conditional on $\boldsymbol{\omega}_{gl}$, $\mathbf{b}_i$, and $\mathbf{a}_i$. According to Equation 11, the sufficient statistic for a_{ij} is $w_{glj} x_{gi}$, and then

$$\frac{\partial A}{\partial a_{ij}} = E\left(w_{glj} x_{gi}\right) = w_{glj} E\left(x_{gi}\right) = w_{glj} \sum_{m'=1}^{M_i} s_{im'} p\left(x_{gi} = s_{im'} | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right)$$

$$\frac{\partial^2 A}{\partial a_{ij} \partial a_{ij'}} = \operatorname{cov}\left(w_{glj} x_{gi}, w_{glj'} x_{gi}\right) = w_{glj} w_{glj'} E\left(x_{gi}^2\right) - w_{glj} w_{glj'} E^2\left(x_{gi}\right)$$

$$= w_{glj} w_{glj'} \sum_{m'=1}^{M_i} s_{im'}^2 p\left(x_{gi} = s_{im'} | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right) - w_{glj} w_{glj'} \left[\sum_{m'=1}^{M_i} s_{im'} p\left(x_{gi} = s_{im'} | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right)\right]^2.$$

The sufficient statistic for b_{it} is $I(x_{gi} = t)$, and then

$$\frac{\partial A}{\partial b_{it}} = E\left[I\left(x_{gi} = t\right)\right] = p\left(x_{gi} = t | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right),$$

$$\frac{\partial^2 A}{\partial b_{it} \partial b_{it'}} = \operatorname{cov}\left[I\left(x_{gi} = t\right), I\left(x_{gi} = t'\right)\right] = \begin{cases} -p\left(x_{gi} = t | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right) p\left(x_{gi} = t' | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right), t \neq t', \\ p\left(x_{gi} = t | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right) - \left[p\left(x_{gi} = t | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right)\right]^2, t = t', \end{cases}$$

$$\frac{\partial^2 A}{\partial a_{ij} \partial b_{it}} = \operatorname{cov}\left[x_{gi} w_{glj}, I\left(x_{gi} = t\right)\right]$$

$$= w_{glj} p\left(x_{gi} = t | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right) \left[t - \sum_{m'=1}^{M_i} s_{im'} p\left(x_{gi} = s_{im'} | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right)\right].$$

Graded Response Function

For the MGRM, the item response function is an extension of Samejima's (1969) unidimensional GRM and can be written as (Cai, 2010a)

$$p_{glis_m} = p\left(x_{gi} = s_{im} | \boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right) = p\left(x_{gi} = s_{im} | \boldsymbol{\omega}_{gl}, \mathbf{b}_i, \mathbf{a}_i\right)$$

$$= p^*\left(x_{gi} = s_{i(m-1)} | \boldsymbol{\omega}_{gl}, b_{is_{i(m-1)}}, \mathbf{a}_i\right) - p^*\left(x_{gi} = s_{im} | \boldsymbol{\omega}_{gl}, b_{is_{im}}, \mathbf{a}_i\right)$$

$$= p^*_{gl\,is_{i(m-1)}} - p^*_{gl\,is_m},$$

where $p^*_{glis_m}$ is the cumulative item response function for $x_{gi} > s_{im}$ conditional on $\boldsymbol{\omega}_{gl}$, $b_{is_{im}}$, and $\mathbf{a}_i$, that is,

$$p^*_{glis_{im}} = p^*\left(x_{gi} = s_{im} | \boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right) = p\left(x_{gi} > s_{im} | \boldsymbol{\omega}_{gl}, b_{is_{im}}, \mathbf{a}_i\right)$$

$$= \frac{\exp\left(b_{is_{im}} + \sum_{j \in \mathrm{J}_g} a_{ij} q_{ij} w_{glj}\right)}{1 + \exp\left(b_{is_{im}} + \sum_{j \in \mathrm{J}_g} a_{ij} q_{ij} w_{glj}\right)},$$

$$p^*_{glis_{i0}} = 1,$$

$$p^*_{glis_{iM_i}} = 0.$$

The first derivative of $\log\left(p_{gis_{im}}\right)$ with respect to item parameter, η_{iv}, is written as

$$\frac{\partial \log\left(p_{glis_{im}}\right)}{\partial \eta_{iv}} = \frac{\left[p^*_{glis_{i(m-1)}} \frac{\partial \log\left(p^*_{glis_{i(m-1)}}\right)}{\partial \eta_{iv}} - p^*_{glis_{im}} \frac{\partial \log\left(p^*_{glis_{im}}\right)}{\partial \eta_{iv}}\right]}{p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}}.$$

Because $p^*_{gis_{i(m-1)}}$ and $p^*_{gis_{im}}$ are just the special cases of the GPCM function for dichotomous items, all the derivative results presented in the preceding section apply to $p^*_{gis_{i(m-1)}}$ and $p^*_{gis_{im}}$. In particular, for $m \neq 0$ or M_i, and $q_{ij} = 1$,

$$\frac{\partial \log\left(p^*_{glis_{im}}\right)}{\partial a_{ij}} = w_{glj} - w_{glj} p^*_{glis_{im}},$$

and for $m \neq 0$ or M_i,

$$\frac{\partial \log\left(p^*_{glis_{im}}\right)}{\partial b_{it}} = \begin{cases} 1 - p^*_{glis_{im}}, & t = s_{im}, \\ 0, & t \neq s_{im}. \end{cases}$$

Then,

$$\frac{\partial \log\left(p_{glis_{im}}\right)}{\partial a_{ij}} = w_{glj}\left(1 - p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}\right),$$

and

$$\frac{\partial \log\left(p_{glis_{im}}\right)}{\partial b_{it}} = \begin{cases} \frac{p^*_{glis_{im}}\left(p^*_{glis_{im}} - 1\right)}{p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}}, & t = s_{im}, \\ \frac{p^*_{glis_{i(m-1)}}\left(1 - p^*_{glis_{i(m-1)}}\right)}{p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}}, & t = s_{i(m-1)}, \\ 0, & \text{otherwise.} \end{cases}$$

For the second derivatives of $\log\left(p_{glis_{im}}\right)$ with respect to item parameters where $m \neq 0$ or M_i,

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial b_{is_{im}} \partial b_{is_{im}}} = \frac{p^*_{glis_{im}}\left(1 - p^*_{glis_{im}}\right)\left(2p^*_{glis_{im}} - 1\right)\left(p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}\right) - p^{*2}_{glis_{im}}\left(p^*_{glis_{im}} - 1\right)^2}{\left(p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}\right)^2},$$

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial b_{is_{i(m-1)}} \partial b_{is_{i(m-1)}}} = \frac{p^*_{glis_{i(m-1)}}\left(p^*_{glis_{i(m-1)}} - 1\right)\left(2p^*_{glis_{i(m-1)}} - 1\right)\left(p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}\right) - p^{*2}_{glis_{i(m-1)}}\left(p^*_{glis_{i(m-1)}} - 1\right)^2}{\left(p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}\right)^2},$$

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial b_{is_{im}} \partial b_{is_{i(m-1)}}} = \frac{p^*_{glis_{i(m-1)}}\left(p^*_{glis_{i(m-1)}} - 1\right) p^*_{glis_{im}}\left(p^*_{glis_{im}} - 1\right)}{\left(p^*_{glis_{i(m-1)}} - p^*_{glis_{im}}\right)^2},$$

and for $t \neq s_{im}$ or $s_{i(m-1)}$, or $t' \neq s_{im}$ or $s_{i(m-1)}$,

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial b_{it} \partial b_{it'}} = 0;$$

if $q_{ij} = 1$ and $q_{ij'} = 1$,

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial a_{ij}\partial a_{ij'}} = -w_{glj}w_{glj'}\left(p^*_{glis_{im}} + p^*_{glis_{i(m-1)}} - p^{*2}_{glis_{im}} - p^{*2}_{glis_{i(m-1)}}\right);$$

if $q_{ij} = 1$,

$$\frac{\partial^2 \log\left(p_{glis_{im}}\right)}{\partial a_{ij}\partial b_{it}} = \begin{cases} w_{glj}p^*_{glis_{im}}\left(p^*_{glis_{im}} - 1\right), & t = s_{im}, \\ w_{glj}p^*_{glis_{i(m-1)}}\left(p^*_{glis_{i(m-1)}} - 1\right), & t = s_{i(m-1)}, \\ 0, & \text{otherwise.} \end{cases}$$

Three-Parameter Logistic Function

The M3PL function for dichotomous items is written as (Reckase, 1997)

$$\begin{aligned} p_{glis_{i2}} &= p\left(x_{gi} = s_{i2}|\boldsymbol{\omega}_{gl}, \boldsymbol{\eta}_{gi}\right) = p\left(x_{gi} = s_{i2}|\boldsymbol{\omega}_{gl}, b_i, \mathbf{a}_i, c_i\right) \\ &= c_i + \left(1 - c_i\right)\frac{\exp\left(b_{is_{i2}} + \sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}\right)}{1 + \exp\left(b_{is_{i2}} + \sum_{j\in \mathrm{J}_g} a_{ij}q_{ij}w_{glj}\right)}, \end{aligned}$$

$$p_{glis_{i1}} = 1 - p_{glis_{i2}},$$

where c_i is the guessing parameter. For easy estimation, define

$$c_i = \frac{\exp\left(c'_i\right)}{1 + \exp\left(c'_i\right)},$$

and c'_i is the item parameter to be estimated. Let

$$W = \frac{\exp\left(b_{is_{i2}} + \sum_{j\in J_{gf}} a_{ij}q_{ij}w_{glj}\right)}{1 + \exp\left(b_{is_{i2}} + \sum_{j\in J_{gf}} a_{ij}q_{ij}w_{glj}\right)}.$$

The derivatives of $\log\left(p_{glis_{i2}}\right)$ with respect to item parameters are shown as follows:
For c'_i

$$\frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial c'_i} = \frac{(1 - W)\exp\left(c'_i\right)}{p_{glis_{i2}}\left[1 + \exp\left(c'_i\right)\right]^2}$$

$$\frac{\partial^2 \log\left(p_{glis_{i2}}\right)}{\partial c'^2_i} = \frac{(1 - W)\,p_{glis_{i2}}\left[1 - \exp\left(2c'_i\right)\right]\exp\left(c'_i\right) - (1 - W)^2\exp\left(2c'_i\right)}{p^2_{glis_{i2}}\left[1 + \exp\left(c'_i\right)\right]^4}.$$

For η_{iv} and $\eta_{iv'}$ in $b_{is_{i2}}$ and $\mathbf{a}_i$

$$\frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial \eta_{iv}} = \left(1 - c_i\right)W\frac{\partial \log(W)}{\partial \eta_{iv}}/p_{glis_{i2}}.$$

Let

$$W_1 = \left(1 - c_i\right)W\frac{\partial \log(W)}{\partial \eta_{iv}}.$$

Then,

$$\frac{\partial W_1}{\partial \eta_{iv'}} = \left(1 - c_i\right)\left[W\frac{\partial \log(W)}{\partial \eta_{iv'}}\frac{\partial \log(W)}{\partial \eta_{iv}} + W\frac{\partial^2 \log(W)}{\partial \eta_{iv}\partial \eta_{iv'}}\right],$$

$$\frac{\partial^2 \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv}\partial\eta_{iv'}} = \left[\frac{\partial W_1}{\partial\eta_{iv'}} - W_1\frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv'}}\right]/p_{glis_{i2}},$$

$$\frac{\partial^2 \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv}\partial c_i'} = \frac{\left[-W\frac{\partial \log(W)}{\partial\eta_{iv}} - (1-W)\frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv}}\right]\exp\left(c_i'\right)}{p_{glis_{i2}}\left[1+\exp\left(c_i'\right)\right]^2}.$$

Again, the derivatives of $\log(W)$ are the special cases of the GPCM function for dichotomous items.

As for $\log\left(p_{glis_{i1}}\right)$, the derivatives with respect to any item parameters η_{iv} and $\eta_{iv'}$ in c_i', $b_{is_{i2}}$, and $\mathbf{a}_i$ are

$$\frac{\partial \log\left(p_{glis_{i1}}\right)}{\partial\eta_{iv}} = \frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv}}/\left(p_{glis_{i2}}-1\right)$$

$$\frac{\partial^2 \log\left(p_{glis_{i1}}\right)}{\partial\eta_{iv}\partial\eta_{iv'}} = \left[\frac{\partial^2 \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv}\partial\eta_{iv'}}\left(p_{glis_{i2}}-1\right) - p_{glis_{i2}}\frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv'}}\frac{\partial \log\left(p_{glis_{i2}}\right)}{\partial\eta_{iv}}\right]/\left(p_{glis_{i2}}-1\right)^2.$$

Attribute Distribution Parameter Estimation

For a multivariate normal distribution vector $\boldsymbol{\omega}_{gl}$, its distribution parameters, mean vector $\boldsymbol{\mu}_g$ with elements μ_{gj}, and variance–covariance matrix $\boldsymbol{\Sigma}_g$with elements $\sigma_{gkk'}$ can be estimated, where k ($k = 1$ to J_g, $k' = 1$ to J_g) represents the kth attribute in group g. Specifically, the estimation is to find $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ maximizing the approximate marginal log-likelihood function given item parameter estimates in $\boldsymbol{\eta}_g$ obtained from the maximization step, which in turn is to maximize the sum (over all test takers in group g) of the expected logarithm of the density of $\boldsymbol{\omega}_{gl}$ with respect to the posterior distribution of $\boldsymbol{\omega}_{gl}$ conditional on $\mathbf{X}_{gp}$, g, and $\boldsymbol{\eta}_g$ (Glas et al., 2000; Li, Bolt, & Fu, 2006), that is,

$$D\left(\boldsymbol{\omega}_{gl}\right) = \sum_{p=1}^{P_g} N_{gp}E\left\{\log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]|\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g\right\},$$

where $\phi(\boldsymbol{\omega}_{gl})$is the multivariate normal density function of $\boldsymbol{\omega}_{gl}$.

Based on the integration rule, *differentiation under the integral sign*, the first derivatives of $D(\boldsymbol{\omega}_{gl})$ with respect to the distribution parameter vector $\boldsymbol{\delta}_g$, where $\boldsymbol{\delta}_g$ contains all parameters in $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, become

$$\frac{\partial D\left(\boldsymbol{\omega}_{gl}\right)}{\partial\,\boldsymbol{\delta}_g} = \sum_{p=1}^{P_g} N_{gp}E\left\{\frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial\boldsymbol{\delta}_g}|\mathbf{X}_{gp}, g, \boldsymbol{\eta}_g\right\}. \tag{12}$$

The derivatives of $\log[\phi(\boldsymbol{\omega}_{gl})]$ with respect to $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ (Anderson, 1958) are

$$\frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial\boldsymbol{\mu}_g} = \boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{\omega}_{gl}-\boldsymbol{\mu}_g\right)$$

$$\frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial\boldsymbol{\Sigma}_g} = -\frac{1}{2}\left[\boldsymbol{\Sigma}_g^{-1} - \boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{\omega}_{gl}-\boldsymbol{\mu}_g\right)\left(\boldsymbol{\omega}_{gl}-\boldsymbol{\mu}_g\right)'\boldsymbol{\Sigma}_g^{-1}\right],$$

where

$$\frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \boldsymbol{\mu}_g} = \left\{\frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \mu_{g1}} \quad \cdots \quad \frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \mu_{gJ_g}}\right\}'; \frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \boldsymbol{\Sigma}_g} = \left\{\begin{matrix} \frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \sigma_{g11}} & \cdots & \frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \sigma_{g1J_g}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \sigma_{gJ_g1}} & \cdots & \frac{\partial \log\left[\phi\left(\boldsymbol{\omega}_{gl}\right)\right]}{\partial \sigma_{gJ_gJ_g}} \end{matrix}\right\};$$

$(\cdot)'$ denotes the transpose of a vector or matrix; and both $\boldsymbol{\omega}_{gl}$ and $\boldsymbol{\mu}_g$ are $J_g \times 1$ vectors. Substitute these two equations into the first derivatives of $D(\boldsymbol{\omega}_{gl})$ (Equation 12) with respect to $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, respectively, and set the derivatives to zero. The solutions of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ from these functions that maximize $D(\boldsymbol{\omega}_{gl})$ (Rayner, 1985) are then

$$\boldsymbol{\mu}_g = \sum_{l=1}^{L_g} \boldsymbol{\omega}_{gl} p\left(\boldsymbol{\omega}_{gl} | \mathrm{X}_{gp}, g, \boldsymbol{\eta}_g\right)$$

$$\boldsymbol{\Sigma}_g = \sum_{l=1}^{L_g} \left(\boldsymbol{\omega}_{gl} - \boldsymbol{\mu}_g\right)\left(\boldsymbol{\omega}_{gl} - \boldsymbol{\mu}_g\right)' p\left(\boldsymbol{\omega}_{gl} | \mathrm{X}_{gp}, g, \boldsymbol{\eta}_g\right).$$

They are just the mean vector and variance–covariance matrix of the posterior distribution of the attributes at the current iteration.

Attribute Estimation

With the estimates of item parameters and continuous attribute distribution parameters available, test takers' attribute vectors can then be estimated. Two approaches for attribute estimation are described in this section: maximum a posterior (MAP) and expected a posterior (EAP).

Maximum A Posterior

The MAP estimate of an attribute vector ($\boldsymbol{\theta}_{gp}$) and/or a latent group membership to a response pattern is just the attribute value (or quadrature point) vector and/or latent group membership having the largest posterior joint probability conditional on the response pattern and item and/or population parameters that are the by-product of the E step (Baker & Kim, 2004). In particular, the MAP estimate of $\boldsymbol{\theta}_{gp}$ is the $\boldsymbol{\omega}_{gl}$ that has maximum $p\left(\boldsymbol{\omega}_{gl} | \mathrm{X}_{gp}, g, \hat{\boldsymbol{\eta}}_g\right)$, and the MAP estimate of the latent group membership is the g that has maximum $\sum_l^{L_g} p\left(\boldsymbol{\omega}_{gl}, g | \mathrm{X}_p, \hat{\boldsymbol{\eta}}_g\right)$.

Expected A Posterior

The EAP estimate of an attribute vector for a response pattern conditional on a group is the expected value of an attribute vector with respect to its joint posterior distribution conditional on the response pattern, group membership, and item and/or population parameters (Baker & Kim, 2004); that is,

$$E\left(\hat{\boldsymbol{\theta}}_{gp} \mid \mathrm{X}_{gp}, \hat{\boldsymbol{\eta}}_g, g\right) = \sum_{l=1}^{L_g} \boldsymbol{\omega}_{gl} p\left(\boldsymbol{\omega}_{gl} \mid \mathrm{X}_{gp}, \hat{\boldsymbol{\eta}}_g, g\right).$$

Variance and Covariance Matrix of Attributes

The covariance matrix of $\boldsymbol{\theta}_{gp}$ estimates for response pattern p in group g (Haberman, von Davier, & Lee, 2008) is

$$\mathrm{cov}\left(\hat{\boldsymbol{\theta}}_{gp}\right) = \sum_{l=1}^{L_g} p\left(\boldsymbol{\omega}_{gl} \mid \mathrm{X}_{gp}, \hat{\boldsymbol{\eta}}_g, g\right)\left(\boldsymbol{\omega}_{gl} - \hat{\boldsymbol{\theta}}_{gp}\right)\left(\boldsymbol{\omega}_{gl} - \hat{\boldsymbol{\theta}}_{gp}\right)'.$$

The covariance matrix of $\boldsymbol{\theta}_g$ estimates for observed group g (Haberman et al., 2008) is

$$\operatorname{cov}\left(\widehat{\boldsymbol{\theta}}_g\right) = \frac{1}{N_g}\sum_{p=1}^{P_g} N_{gp}\left(\widehat{\boldsymbol{\theta}}_{gp} - \overline{\boldsymbol{\theta}}_g\right)\left(\widehat{\boldsymbol{\theta}}_{gp} - \overline{\boldsymbol{\theta}}_g\right)',$$

where

$$\overline{\boldsymbol{\theta}}_g = \frac{1}{N_g}\sum_{p=1}^{P_g} N_{gp}\widehat{\boldsymbol{\theta}}_{gp}.$$

For latent group g,

$$\operatorname{cov}\left(\widehat{\boldsymbol{\theta}}_g\right) = \frac{1}{\sum_{l=1}^{L_g}\overline{n}_{gl}}\sum_{p=1}^{P_g}\sum_{l=1}^{L_g} N_p p\left(\boldsymbol{\omega}_{gl}, g|\mathbf{X}_p, \widehat{\boldsymbol{\eta}}_g\right)\left(\widehat{\boldsymbol{\theta}}_{gp} - \overline{\boldsymbol{\theta}}_g\right)\left(\widehat{\boldsymbol{\theta}}_{gp} - \overline{\boldsymbol{\theta}}_g\right)',$$

where

$$\boldsymbol{\theta}_g = \frac{1}{\sum_{l=1}^{L_g}\overline{n}_{gl}}\sum_{p=1}^{P_g}\sum_{l=1}^{L_g} N_p\widehat{\boldsymbol{\theta}}_{gp} p\left(\boldsymbol{\omega}_{gl}, g|\mathbf{X}_p, \widehat{\boldsymbol{\eta}}_g\right).$$

Simulation Study: Model Parameter Recovery

The MML-EM estimation procedure described previously was implemented in the freeware program R. A simulation study was conducted to examine item and/or attribute distribution parameter recovery on nine GPCM-related models. These models varied by the number of item parameters (one or two), number of item score categories (two or three), number of attributes (one or two), number of groups (one or two), and group feature (observed or latent). The basic models are GPCM and its two submodels for dichotomous items, 1PLs and 2PLs. The tradition 1PL, 2PL, and GPCM models are just the one-group one-attribute models. Following are some key points on the setup of the simulation study:

1. A simulated data set included 10 items, either all dichotomous items or three-category polytomous items. Dichotomous items were scored as 0 or 1, and polytomous items were scored as 0, 1, or 2. Discrimination parameters were drawn from a lognormal distribution with mean 1 and variance .04 on the normal scale. For models with two attributes, some items only required one attribute, while the others required both. Intercept parameters were drawn from the standard normal distribution. For the multigroup models with two observed groups, the item parameters of all items, except one, were constrained to be equal across the two groups. For the mixture models with two latent groups, no item parameter was constrained to be equal across the two groups.
2. A simulated data set included 3,000 test takers. For two-group (multigroup or mixture) models, test takers were evenly divided into groups. For one-group models with one attribute, the continuous attribute was sampled from the standard normal distribution. For one-group models with two attributes, the two continuous attributes were sampled from the standard bivariate normal distribution with a correlation of .7. For two-group models with one attribute, the attribute for one group was sampled from the standard normal distribution and for the other group was sampled from the normal distribution with mean 1 and standard deviation 1. For two-group models with two attributes, the attribute for one group was sampled from the standard bivariate normal distribution with a correlation of .7 and for the other group was sampled from the bivariate normal distribution with mean vector (1, 1) and correlation .5.
3. Model identification was achieved by imposing the following constraints during estimation. For one-group models with one attribute, the mean and standard deviation of the prior distribution of the attribute were fixed to 0 and 1, respectively. For one-group models with two attributes, the means and standard deviations of the prior distribution of the two attributes were fixed to 0 and 1, respectively; the correlation was freely estimated. For multigroup models with two groups and one attribute, the mean and standard deviation of the prior distribution of the attribute for the first group were fixed to 0 and 1, respectively, and all item parameters, except for those of one item, were constrained to be equal across the two groups. The mean and standard deviation of the prior distribution of the attribute for the second group were freely estimated. For multigroup models with two groups and two attributes, the means and standard deviations of the prior distribution of the two attributes for the first group were fixed to 0 and 1,

Table 1 Root Mean Square Error of Model Parameter Estimates

Model	Discrimination	Intercept	Attribute population mean	Attribute population standard deviation and/or correlation
1PL model	NA	.047	NA	NA
2PL model	.078	.057	NA	NA
GPCM model	.052	.076	NA	NA
Two-dimensional GPCM model	.102	.083	NA	.051
Two-group GPCM model	.084	.088	.161	.064
Two-group two-dimension GPCM model	.114	.094	.067	.051
Two-group mixture 1PL model	NA	.094	NA	NA
Two-group mixture GPCM model	.091	.117	NA	NA
Two-group two-dimensional mixture GPCM model	.177	.130	NA	.058

Note. 1PL = one-parameter logistic. 2PL = two-parameter logistic. GPCM = generalized partial credit model. NA = not applicable.

respectively, and all item parameters, except for those of one item, were constrained to be equal across the two groups. The mean vector and variance–covariance matrix of the prior distribution of the two attributes for the second group as well as the attribute correlation for the first group were freely estimated. For mixture models with two groups, all the parameters of the prior distribution of attribute(s) were fixed to their true values, except that the correlations between the two attributes in both groups were freely estimated. For models with one attribute, the attribute distribution was approximated by 40 quadrature points; for two attributes, the attribute distribution was approximated by a 30×30 quadrature grid.

4. The model convergence criterion during estimation was that both the largest change of item parameter estimate and the change of the approximate observed log-likelihood, $-2\log\left(L_{\text{obs}}^{\text{appr}}\right)$, between two consecutive iterations be smaller than .001. For the M step, only one iteration was allowed. For each model, 30 data sets were generated and estimated. The evaluation criterion for parameter recovery was the root mean square error (RMSE) of parameter estimates across the 30 data sets.

Table 1 lists the RMSEs of model parameter estimates for each model separated into four categories: item discrimination, item intercept, mean, and standard deviation/correlation of the population distribution of attribute(s). From Table 1, one can see that (a) when the models became more complicated (i.e., including more groups and/or attributes), their parameter recovery became somewhat worse, and that (b) parameter recovery was more difficult for the mixture models than the multigroup models. Overall, parameter recovery was reasonable for these models.

Summary

The extended MML-EM provides a nicely integrative framework to estimate all commonly used IRT models (i.e., GPCM, GRM, and 3PL) with multiple (observed or latent) groups and multiple attributes. All estimation functions and derivatives are provided in this report. This procedure was implemented in an R program. A simulation study was conducted using this R program and showed reasonable parameter recovery.

The inverse of the negative Hessian matrix was used to approximate the variance–covariance matrix of model parameter estimates. However, this estimation of standard errors of model parameter estimates has been shown to have bias because the Hessian matrix is not based on the complete data (Liu, Xin, Andersson, & Tian, 2019; Sundberg, 1974). Many estimation methods for standard errors of model parameter estimates have been proposed and compared; see Liu et al. (2019) for a summary and further discussion. Those standard error estimation methods are easy to incorporate into the extended MML-EM estimation procedure developed in this report.

The quadrature space grows exponentially with the number of attributes. Based on the author's experience, each attribute should have at least 20 quadrature points for models with known attribute distribution parameters and at least 30 quadrature points for models with estimated attribute distribution parameters to achieve adequate accuracy of parameter estimation. Thus, for models with more than three attributes, the excessive number of total quadrature points becomes a serious burden for computation. This issue has been addressed by using adaptive quadrature (Haberman, 2006; Schilling & Bock, 2005), stochastic methods (Cai, 2010b; von Davier & Sinharay, 2007), and dimension reduction

via utilizing the special feature of attribute structure, for example, the bifactor models (Cai, 2010a; Gibbons & Hedeker, 1992) and, more generally, the graphical models (Rijmen, 2009). The models and the estimation procedure presented in this report provide a basis for these advancements.

Acknowledgments

Thanks are due to Shelby Haberman, Xueli Xu, Carolyn Wentzel, James Carlson, and Yanming Jiang for their helpful suggestions on and edits to early versions of this report. The author is grateful to Holly Monteith for her editorial assistance on this report. Any remaining errors are mine alone.

References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York, NY: John Wiley.

Atkinson, K. E. (1989). *An introduction to numerical analysis* (2nd ed.). New York, NY: John Wiley.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781482276725

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. New York, NY: John Wiley.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57. https://doi.org/10.1007/s11336-009-9136-x

Cai, L. (2017). *flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Davis, P. J., & Rabinowitz, P. (1984). *Methods of numerical integration*. Cambridge, MA: Academic Press. https://doi.org/10.1016/B978-0-12-206360-2.50012-1

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38. https://doi.org/10.2307/2984875

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423–436. https://doi.org/10.1007/BF02295430

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement* (Vol. 1, pp. 236–258). Norwood, NJ: Ablex.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Boston, MA: Kluwer. https://doi.org/10.1007/0-306-47531-6_14

Haberman, S., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (Research Report No. RR-06-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02035.x

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Heiss, F., & Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, *144*, 62–80. https://doi.org/10.1016/j.jeconom.2007.12.004

Jaeckel, P. (2005). *A note on multivariate Gauss–Hermite quadrature*. Retrieved from the Semantic Scholar website: https://www.semanticscholar.org/paper/A-note-on-multivariate-Gauss-Hermite-quadrature-Jaeckel/0e3932b776eb3803e0f0ae2e414f69399daa411d

Lang, S. (1997). *Undergraduate analysis* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2698-5

Li, Y., Bolt, D. M., & Fu, J. (2006, April). *A multiple-group testlet model and its application to DIF assessment*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology*, *72*, 18–37. https://doi.org/10.1111/bmsp.12134

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement*, *24*, 211–223. https://doi.org/10.1177/01466210022031679

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90. https://doi.org/10.1177/014662169501900109

R Core Team. (2017). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rayner, J. C. W. (1985). Maximum likelihood estimation of μ and Σ from a multivariate normal distribution. *American Statistician*, *39*, 123–124. https://doi.org/10.1080/00031305.1985.10479410

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6

Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Research Report No. RR-09-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02160.x

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *The British Journal of Mathematical and Statistical Psychology*, *44*, 75–92. https://doi.org/10.1111/j.2044-8317.1991.tb00951.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(Supplement 1), 1–97. https://doi.org/10.1007/BF03372160

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555. https://doi.org/10.1007/s11336-003-1141-x

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, *1*, 49–58. https://doi.org/10.2307/4615553

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement*, *23*, 283–297. https://doi.org/10.1177/01466219922031400

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, *58*, 395–415. https://doi.org/10.1007/978-1-4612-4230-7_10

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics*, *32*, 233–251. https://doi.org/10.3102/1076998607300422

von Davier, M., & Xu, X. (2009). Estimating latent structure models (including diagnostic classification models) with mdltm—A3 software for multidimensional discrete latent trait models [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement*, *28*, 389–406. https://doi.org/10.1177/0146621604268734

Suggested citation

Fu, J. (2019). *Maximum marginal likelihood estimation with an expectation-maximization algorithm for multigroup/mixture multidimensional item response theory models* (Research Report No. RR-19-35). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12272

Action Editor: James Carlson

Reviewers: Shelby Haberman and Xueli Xu